Softiwarp

A Software iWARP Driver for OpenFabrics

Bernard Metzler, Fredy Neeser, Philip Frey
IBM Zurich Research Lab

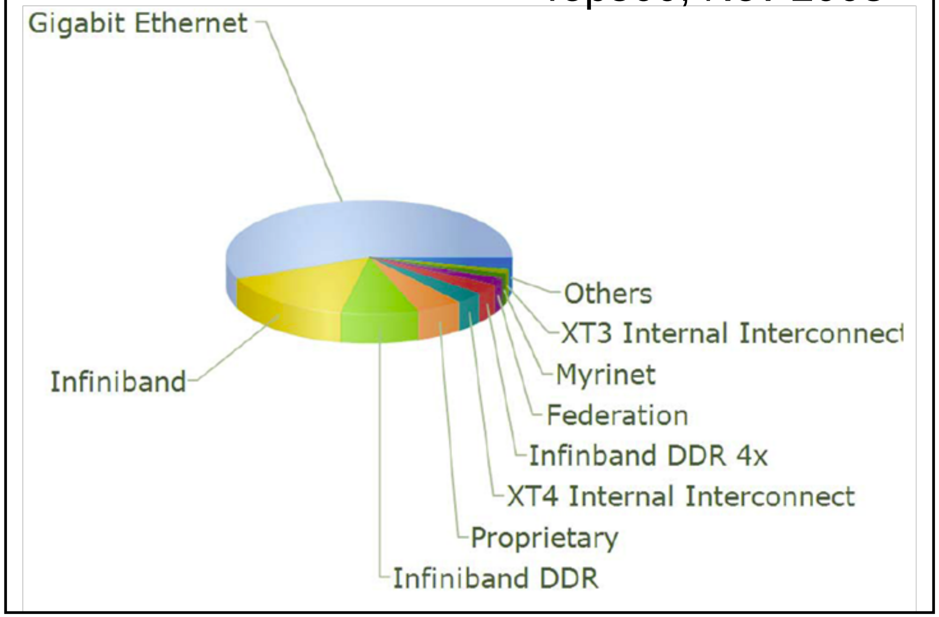
Contents

- Background
- What is it?
- Do we need Software RDMA?
- How is it made?
- Some first Test Results
- Feedback: OFED Issues
- Project Status & Roadmap
- Summary

Background

- RDMA (from proprietary to standard):
  - via, Quadrics, Myrinet, .., InfiniBand, iWARP
- Ethernet (from lame to fast):
  - 1,10,100,1000,10000,40000,…MBit
- Unified Wire:
  - Single link, single switch, single tech. or dump adapter

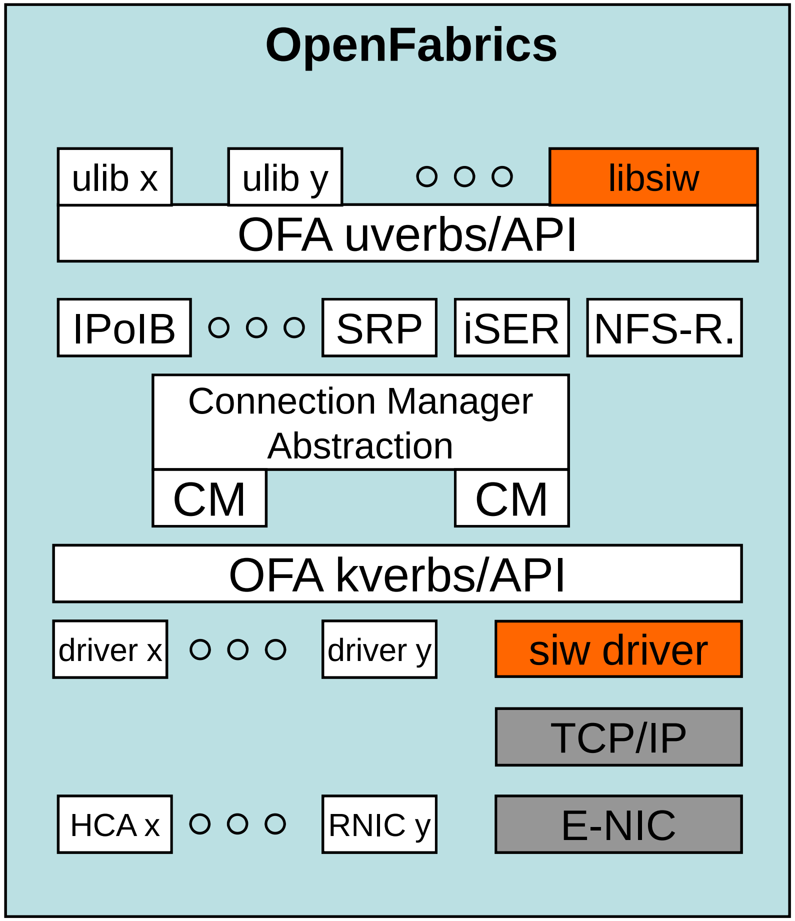
Top500, Nov 2008



- OpenIB
  - Focused on InfiniBand
- OpenFabrics
  - InfiniBand + iWARP HW
  - + iWARP SW?
- IBM Zurich Research
  - RDMA API standardization
  - IETF work on iWARP
  - Software iWARP stack

Softiwarp: What is it?

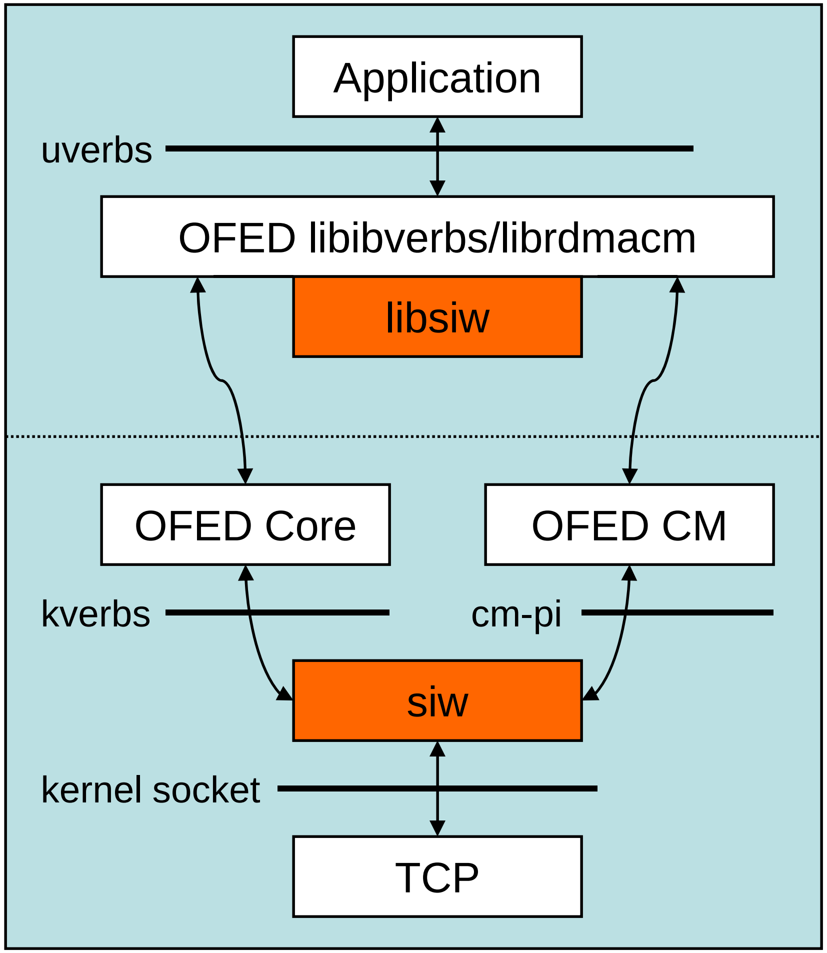
- Just another OFED iWARP driver
  - ../hw/cxgb3/, ../hw/siw,
- Purely software based iWARP protocol stack implementation
  - Kernel module
  - Runs on top of TCP kernel sockets
  - Exports OFED Interfaces (verbs, IWCM, management, …)
- Client support
  - Currently only user level clients
  - libsiw: user space library to integrate with libibverbs, librdmacm
- Current build
  - OFED 1.3
  - Linux 2.6.24
  - ~9000 lines for *.[ch] including comments



OFED and Kernel Integration

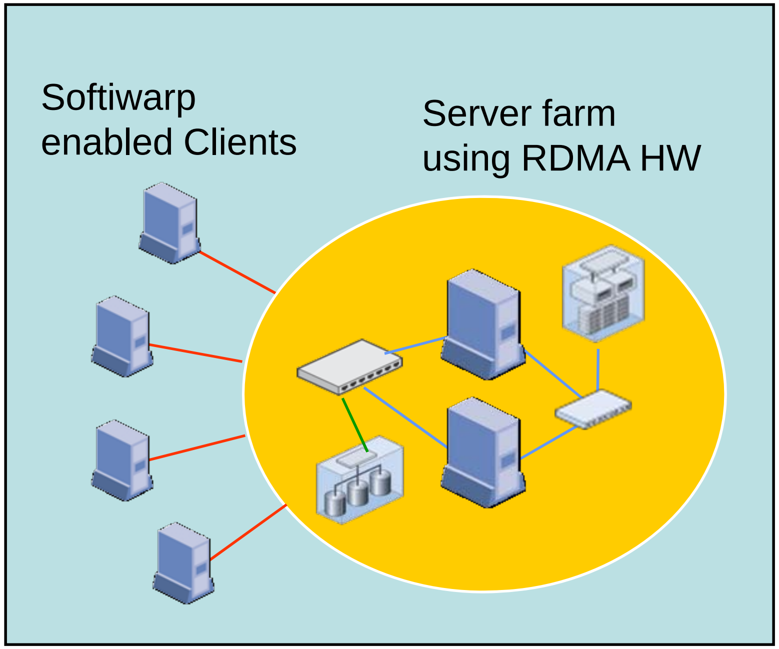
Approach: **Keep things simple and standard**

- TCP interface: Kernel Sockets
  - TCP stack completely untouched
  - Non-blocking write() with pause and resume
  - softirq-based read()
- Linux Kernel Services
  - List-based QP/WQE management
  - Workqueue-based asynchronous sending/CM
  - …
- OFED interface
  - verbs,
  - Event callbacks,
  - Device registration
- Fast Path
  - No private interface between user lib and kernel module
  - Syscall for each post(SQ/RQ) or reap(CQ) operation



Why RDMA in Software?

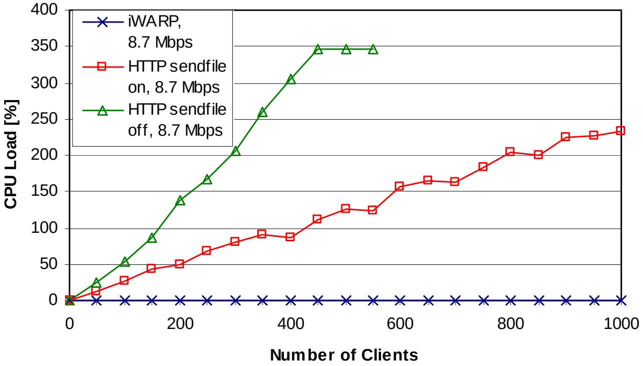
- Enable systems without RNIC to speak RDMA
  - Conventional ENIC sufficient
  - Peer with real RNICs
    - Help busy server to offload
    - Speak RDMA out of the Cluster
    - Enable real RNICs(!)
  - Benefit from RDMA API semantics
    - Application benefits
      - Async. comm., parallelism
      - One-sided operations
    - CPU benefits
      - Copy avoidance in tx
      - Named buffers in rx
- Early system migration to RDMA
  - Migrate applications before RNIC avail.
  - Mix RNIC equipped systems with ENICs
- Test/Debug real HW
- RDMA transport redundancy/failover
- Help to grow OFED Ecosystem for Adoption and Usage beyond HPC

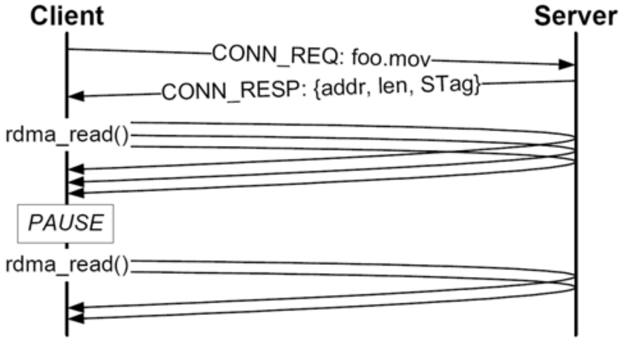


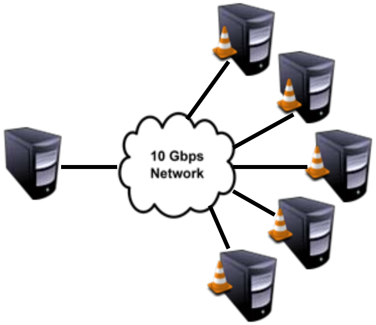
RDMA Use Case != HPC

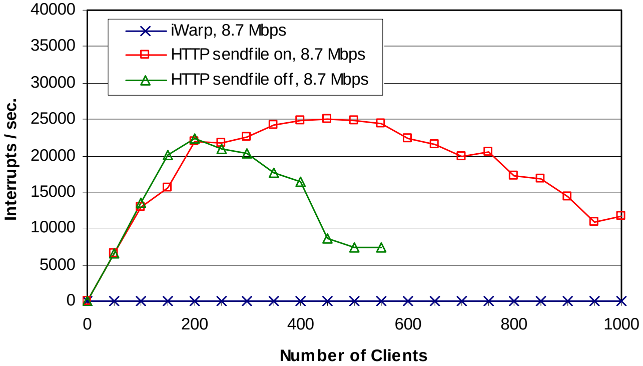
Multimedia Data Dissemination via RDMA

- RNIC-equipped video server, Chelsio t3 10Gb
  - Complete content in Server RAM
  - IBM HS21 BladeServers (4core Xeon 2.33 GHz, 8GB Mem.)
- Up to 1000 VLC clients to pull FullHD (8.7Mbps)
  - VLC client extended for OFED verbs
  - Client may seek in data stream
- HTTP get (Apache w/sendfile()) or RDMA READ
  - Service degradation w/o sendfile
  - Increasing load with sendfile
  - Zero server CPU load for RDMA
- Very simple pull protocol for RDMA
  - Minimum iWARP server state per Client: RDMA READ!



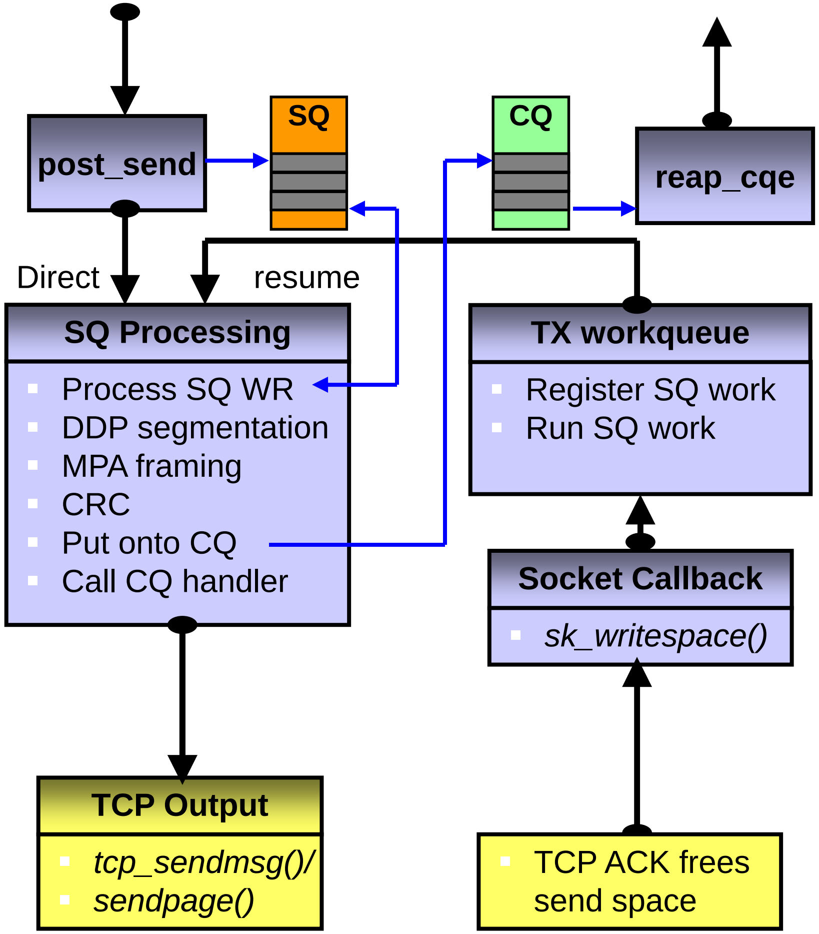






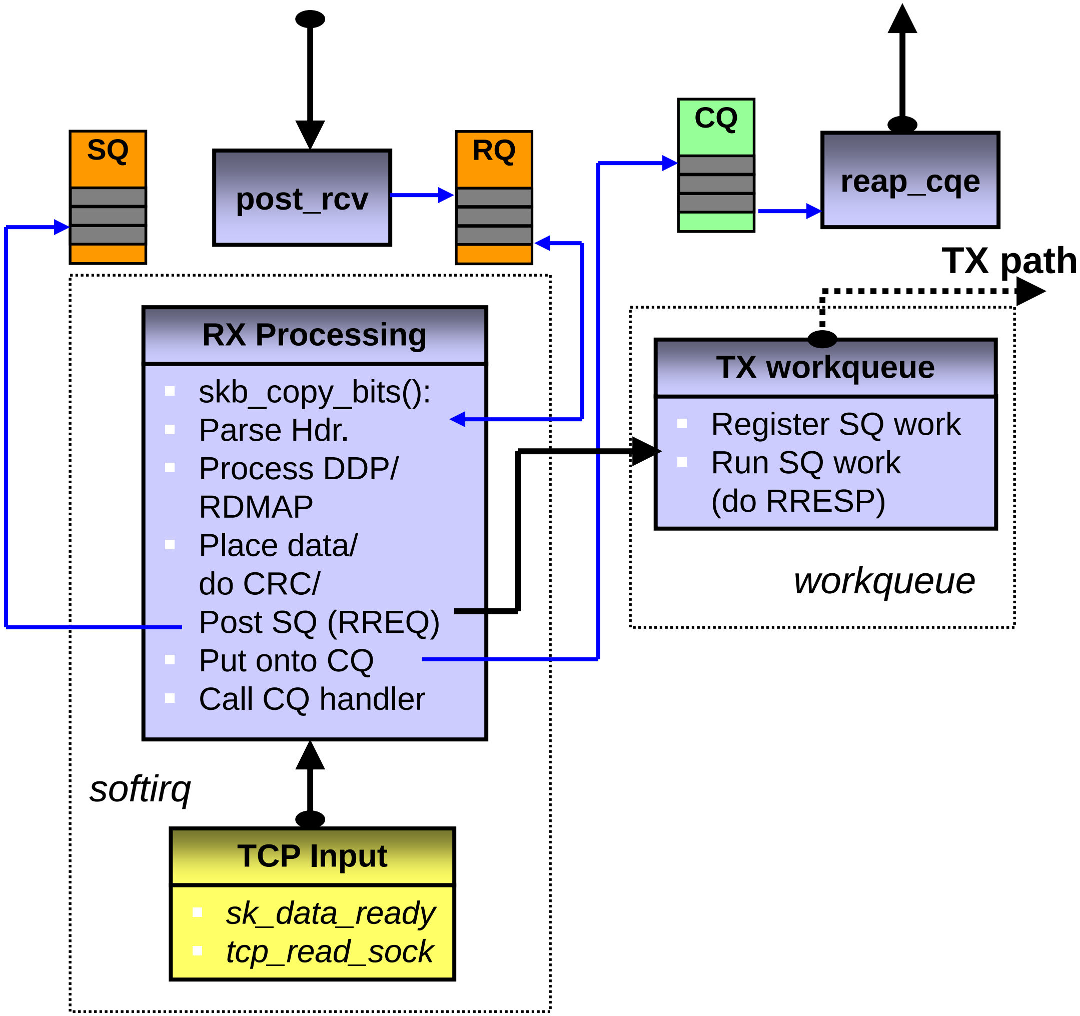
Softiwarp TX Path Design

- Syscall through OFED verbs API to post SQ work
- Synchronous send out of user context if socket send space available
- Nonblocking socket operation:
  - Pause sending if socket buffer full
  - Resume sending if TCP indicates sk_writespace()
    - Use Linux workqueue to resume sending
- Lock-free source memory validation on the fly
- sendfile()-semantic possible
- Post work completions onto CQ
- Reap CQE's asynchronously



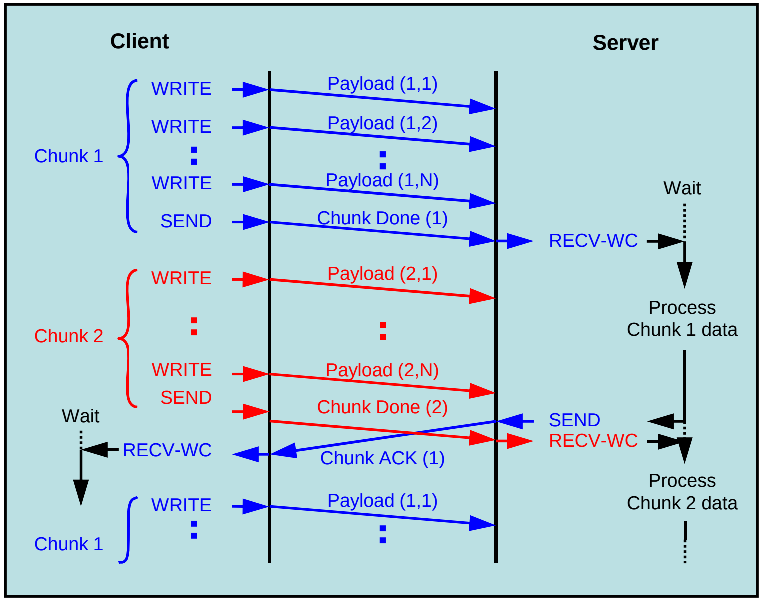
Softiwarp RX Path Design

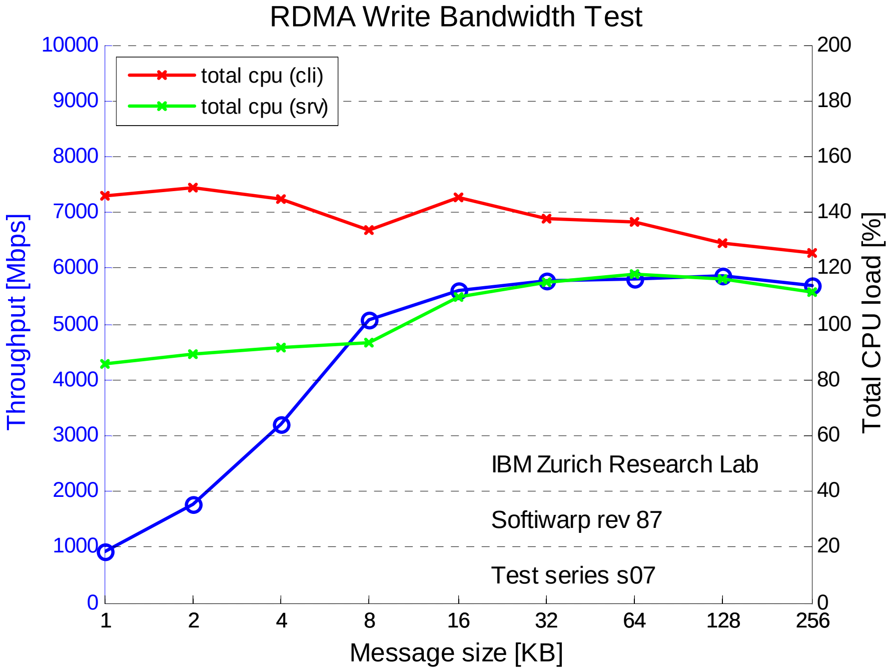
- All RX processing done in softirq context:
  - in sk_data_ready() upcall:
  - Header parsing
  - RQ access
  - Immediate data placement
  - CRC
  - No context switch
  - No extra thread
- Lock-free target memory validation on the fly
- Inbound RREQ just posted at SQ + SQ processing scheduled to resume later



First Tests: Softiwarp

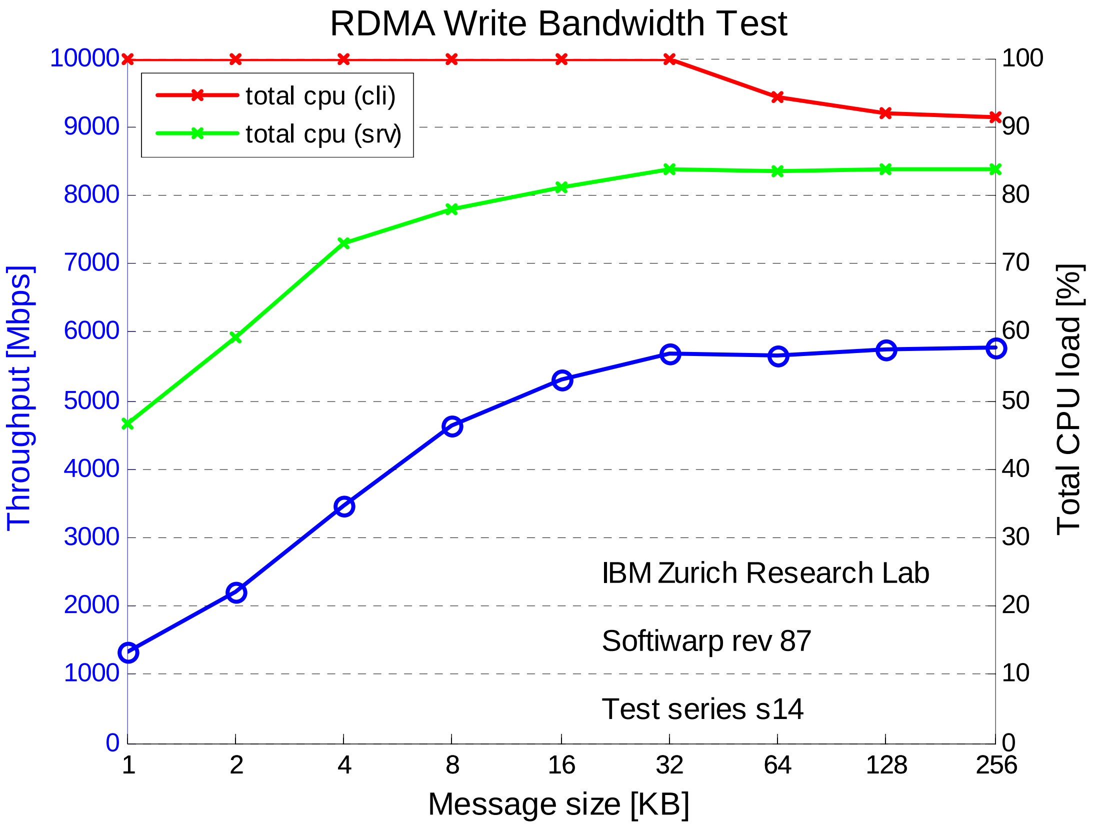
- Non-tuned software stack on both sides
- Application level flow control (ping-pong buffers)
- SEND‘s for synchronization
- 1 connection

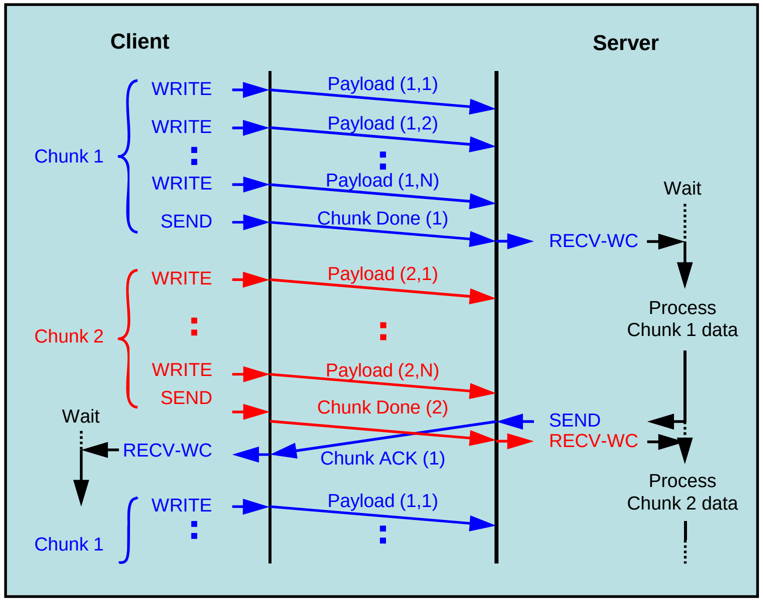




Write/read application data: off
MPA CRC32C: off
MTU = 9000

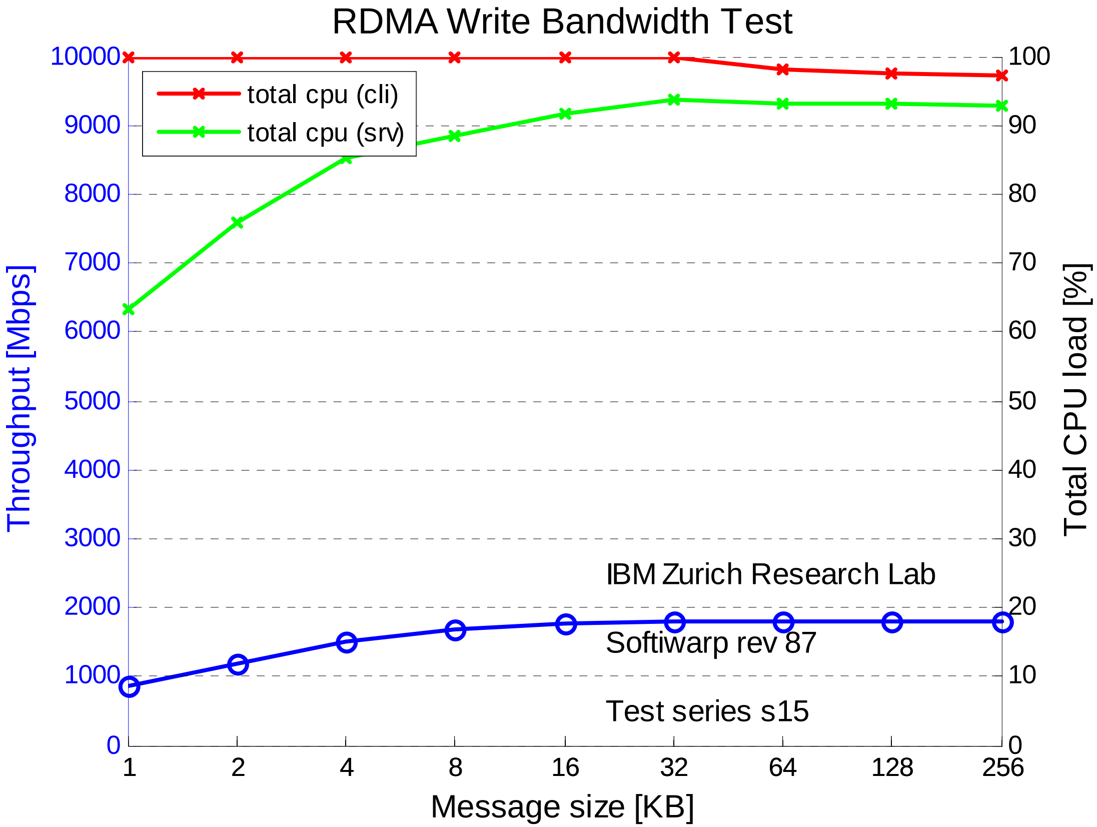
First Tests: Softiwarp

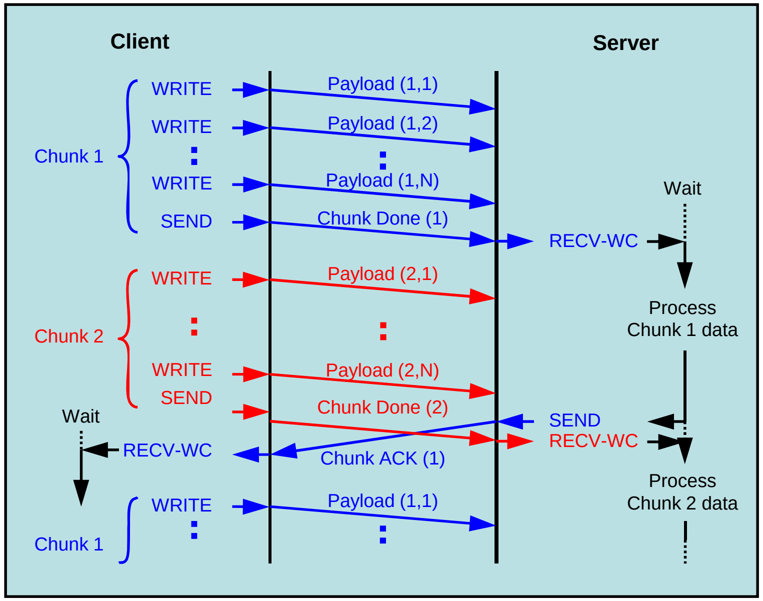




- Same application level flow control (ping-pong buffers) +
  - 1 Core only
  - MPA CRC off
  - MTU=9000
- Sending CPU on its limit

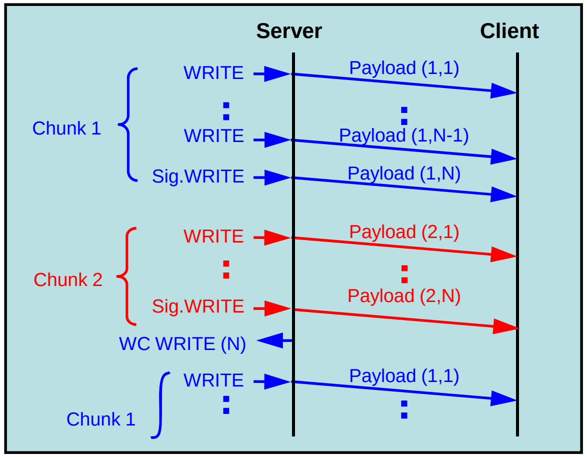
First Tests: Softiwarp + CRC

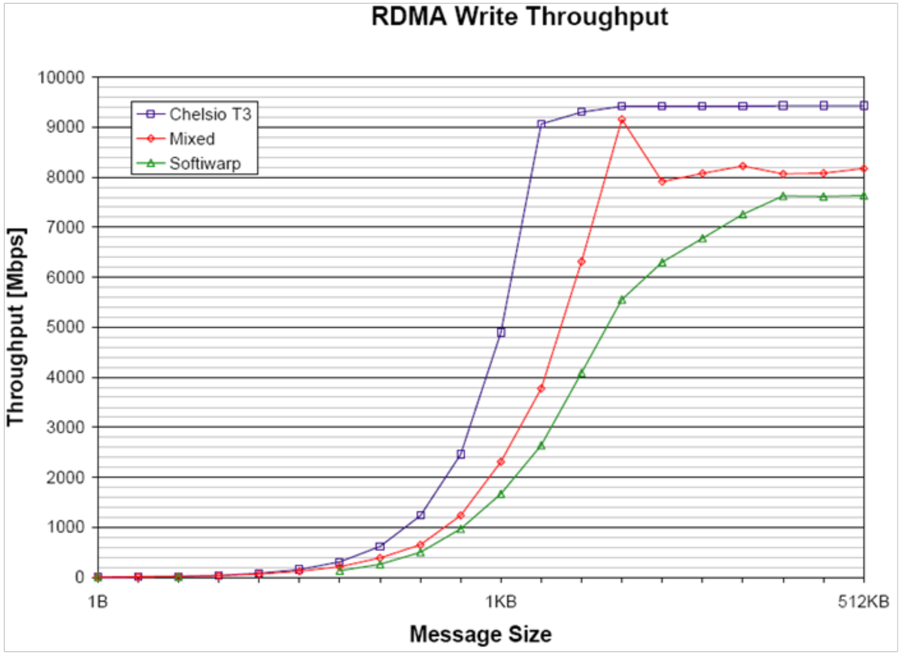




- Same application level flow control (ping-pong buffers) +
  - 1 Core only
  - **MPA CRC ON**
  - MTU=9000
- CRC is killing performance
- Still sending CPU on its limit

First Tests: Softiwarp-Chelsio





- Test 1: Softiwarp peering Chelsio T3
  - Setup:
    - RNIC sends WRITEs to Softiwarp target
    - Target just places data w/o appl. interaction
  - Result:
    - Close to line speed at 8KB
    - Uups - some issues at larger buffers

- Test 2: Softiwarp peering Softiwarp
  - Same setup
  - Result:
    - Maximum Bandwidth from 128KB on

- Conclusions:
  - Promising for first test on non-tuned stack
  - Software stack may perform well on client side
  - Further improvement with sendfile() possible

Softiwarp: Work in progress

Core Functionality

RDMAP/DDP/MPA	x
QP/CQ/PD/MR Objects	x
Send	x
Receive	x
RDMA WRITE	x
RDMA READ	x
Connection Mgmt (IWCM, TCP)	x
Memory Management	x

Features (incomplete)…

MPA CRC	x
MPA Markers	-
Memory Windows	w
Inline Data	w
Shared Receive Queue	-
Fast Memory Registration	-
Termination Messages	w
Remote Invalidation	-
Stag 0	-
Resource Realloc. (MR/QP/CQ)	-
TCP header alignment	w
Relative adressing (ZBVA)	w

(x): done, (w): work in progress, (-) not done

Softiwarp Roadmap

- Opensource very soon
- Discuss current code base in the community
  - Be open for changes/critics
  - Identify core must-haves which are missing
  - Stability!
  - Invite others to contribute
  - Feedback known issues of OFED core to team
  - Don’t touch TCP
- Start compliance testing (OFA IWG) soon
- Investigate private fast path user interface option
- Start working on kernel client support
- Investigate partially offloading of CPU intensive tasks
  - CRC, tx-markers
  - Data placement,..

Feedback: OFED Issues

- Late RDMA Transition
  - Something not part of RNIC integration is now possible
  - Very simple to do with Softiwarp, benefits iSER & Co.
  - Softiwarp allows late RDMA mode transition w/o TCP context migration
- OFED CM
  - How to coexist with RNIC if SW stack shares link, shall we?
  - Can we exist within OFED w/o full (complex) IWCM support?
  - Current code spends 2000 lines out of 9000 for CM!
- Device Management
  - Wildcard listen on multiple interfaces must be translated to individual wildcard listens on each (port/ipaddr) combination
- Zero based virtual adressing
- …

Summary

- Software RDMA is useful
- Software RDMA is efficient on client side (at least)
- RDMA semantics help to use transport efficiently
- Softiwarp helps to grow RDMA/OFED ecosystem
  - Establish RDMA communication model
  - Prepare applications to use RDMA
  - Prepare systems to introduce RDMA HW
  - Peer & thus enable RDMA HW
- Softiwarp is work in progress
  - Please join.